BIOMEDICAL NAME RECOGNITION:

A MACHINE LEARNING APPROACH

ZHANG JIE

(B.Eng., SJTU, PRC)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2003

Name: Zhang Jie
Degree: M.Sc.
Dept: Computer Science, School of Computing
Thesis Title: Biomedical Name Recognition: A Machine Learning Approach

ABSTRACT

The purpose of this research is to automatically recognize names in biomedical documents. First, we analyze characteristics of biomedical names. Then, we propose a rich set of features, including orthographic, morphological, part-of-speech and semantic trigger features. All these features are integrated via a Hidden Markov Model with back-off modeling. Finally, we propose a method for biomedical abbreviation recognition and two methods for cascaded name recognition. Evaluation on GENIA corpus V3.02 and V1.1 shows our system achieves 66.5 and 62.5 F-measure respectively. It shows that our system outperforms previous best published system on the same V1.1 by 8.1 F-measure. The major contribution of this thesis lies in its detailed analysis of biomedical names, the rich feature set and the effective methods for cascaded name recognition. To our best knowledge, our system is the first one that resolves the phenomena of cascaded biomedical names. In addition, a demo has been put on the web.

Keywords: biomedical name recognition, Hidden Markov Model, cascaded name recognition

ACKNOWLEDGEMENTS

I would like to express my great gratitude to my supervisor Dr. Zhou Guodong, my supervisor A/P Tan Chew Lim, and Dr. Su Jian for their advice, guidance and support throughout the duration of my postgraduate study. They have been always accessible and holding discussion and meetings periodically. Their insightful opinions are very important to this thesis.

I also thank the Department of Computer Science, School of Computing, NUS and Institute for Infocomm Research for providing me the opportunity and financial support to study in NUS.

I would like to thank my parents and Miss Shen Dan for their concern, help and support. Without them, I would never be able to fulfill my study. I also thank my lab-mates Mr. Hong Huaqing, Mr. Yang Xiaofeng and other friends for their discussion and help.

TABLE OF CONTENTS

SUMMARY

Biomedical name recognition is to recognize names of entities and concepts in biomedical documents. This task is critical for information extraction and knowledge mining in the biomedical domain. In the last few years, various hand-written rule-based and machine learning approaches have been studied in this field. Among them, most approaches are adapted from the previous MUC named entity recognition task in the newswire domain. However, biomedical names have special characteristics, such as long length distribution, complex naming conventions, various constructions, etc., which are quite different from name entities in the newswire domain. In this thesis, we first analyze these characteristics in the biomedical domain and compare them with the entity names in the newswire domain. Then, we propose various features, such as orthographic, morphological, part-of-speech, and semantics trigger features to deal with the special characteristics in the biomedical names. Especially, we adapt a part-of-speech tagger to the biomedical domain and find that proper adaptation makes a big difference in biomedical name recognition. Finally, we present a method for biomedical abbreviation recognition and two methods for cascaded name recognition which are new to this field. All the features and the methods are integrated via a Hidden Markov Model with back-off modeling. Extensive experiments have been done on the GENIA corpus V3.02 to show the effectiveness of our rich feature set, part-of-speech feature adaptation, and proposed methods on different training data sizes. It shows that the orthographic feature, the morphological feature, the part-of-speech feature adapted into the biomedical domain and the head noun trigger feature are useful for biomedical name recognition, while the special verb trigger feature

cannot lead to positive effect in our model. It also shows that our proposed methods for abbreviation recognition and cascaded name recognition are effective. In addition, our system outperforms previous best published system by 8.1 f-measure on the same GENIA corpus V1.1 without using any dictionary. Finally, a detailed error analysis is done and shows that about 46% of errors come from inconsistent annotation in the corpus. It suggests that a much higher performance can be expected with a more consistent corpus.

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

With the exploding amount of literatures in the biomedical domain, it becomes more and more difficult for people to deal with such a huge amount of text resources. Intelligent techniques must be developed to alleviate human work. As an essential one of these techniques, name entity recognition (NER) automatically identifies names from texts and classifies them into predefined classes. The task of named entity recognition was defined by the Message Understanding Conferences (MUC), which recognizes names of entities such as *PERSON*, *ORGANIZATION, LOCATION* and etc in the newswire domain. In the biomedical domain, not only the names of entities, such as protein, gene and virus, but also the names of some concepts, such as names of biomedical process, are needed to be recognized. Therefore, we use the term "biomedical name recognition" for the named entity recognition task in the biomedical domain so as to be more general. Biomedical name recognition can be widely applied in:

1) Text Mining in the biomedical domain (e.g. protein-protein interactions from literature)
2) Information Extraction
3) Information Retrieval
4) Biomedical Databases (e.g. automatic database building and updating)
5) Question & Answering

Previous research work of named entity recognition in MUC got some promising results in the newswire domain. However, due to special characteristics in the biomedical domain, traditional named entity recognition techniques fail to achieve satisfactory results in biomedical name recognition. The purpose of our work is to study the special characteristics of biomedical names and develop a new name recognition model for the biomedical domain.

1.2 Background

The Message Understanding Conferences (MUC), sponsored by DARPA in the U.S.A., defined the task of named entity recognition. The task consists of three subtasks including recognition of named entity, temporal expression and number expression. Among the three subtasks, the named entity subtask recognizes person names, location names and organization names; the temporal expression subtask recognizes date and time expressions; and the number expression subtask recognizes quantity expressions of monetary values and percentages. Generally speaking, the named entity recognition task can be regarded as a combination of two procedures: entity identification and entity classification. Entity identification tries to find the boundaries of all named entities and entity classification assigns a type for each identified instances.

In the early days of named entity recognition, people mainly relied on the manually written rules and pattern-matching techniques. The LTG system for MUC-7 [Mikheev et al. 1998] uses 5 phase probabilistic partial matching rules for the named entity subtask and a special developed grammar and compiled lists for the temporal and numeric subtasks.

The NetOwl Extractor System [Krupka and Hausman 1998] recognizes named entities based on lexicons and pattern rules. The LaSIE-II system described in [Humphreys et al. 1998] also makes use of hand-coded rules. Although most of these systems get quite high performance, they relied much on manual work and adapting them to a new domain is normally time-consuming and very expensive.

Besides hand-coded rule-based approaches, there are more and more studies using machine-learning techniques for named entity recognition. Many machine learning approaches have been applied in the task of named entity recognition, including Maximum Entropy (ME) [Borthwick et al. 1998; Chieu and Ng 2002], Hidden Markov Models (HMM) [Miller et al. 1998; Bikel et al. 1999; Yu et al. 1998; Zhou and Su 2002], Decision Trees and Support Vector Machines (SVM). Along with them, many research works have been exploring various features for named entity recognition. [Chieu and Ng 2002] made use of the global information. [Zhou and Su 2002] presented an effective constraint relaxation algorithm for solving data sparseness problem in order to integrate rich feature sets. Comparatively speaking, machine learning approaches are more capable of adaptation and cheaper for maintenance.

1.3 Related Work

This section presents a review of recent literatures on biomedical name recognition. With fast development in the research field of biology and life science, there are more and more research projects on biomedical name recognition. Some of them are adapted from the

previous MUC systems. From the methodological point of view, all of them can be grouped into rule-based and machine learning-based.

1.3.1 Rule-based Approaches

As for rule-based approaches, the representative research efforts include [Fukuda et al. 1998], [Proux et al. 1998] and [Gaizauskas et al. 2000].

[Fukuda et al. 1998] proposed a method called PROPER (Protein Proper-noun phrase Extracting Rules), which attempted to identify protein names from biomedical documents based on surface clues of character strings, such as the presence of upper cases and special characters. They summarized the nomenclature of protein names into three categories based on the surface characteristics of word. They defined strings with special orthographic patterns as "core-terms", such as string with capital letters, digits and special symbols. They also defined a list of keywords called "f-terms (feature-terms)" to determine functions or to compose compound words. From our understanding, the "f-terms" are basically the head nouns of compound protein names. The method identifies protein names in two phases. First, it extracts "core-terms" from tokenized texts by five hand-written rules. Second, it concatenates "core-terms" and "f-terms" by rebuilding "core-blocks" and "dependency-blocks". Every "core-block" is a noun phrase without conjunctions and prepositions, which was concatenated by "core-terms" and "f-terms" using a series of rules. Every "dependency-block" is a block between "core-blocks", which basically solves conjunction problem. The evaluation on 30 annotated MEDLINE abstracts on SH3 domain achieved precision of 91.90% and recall of 93.32%.

[Proux et al. 1998] detected gene names in biomedical documents based on lexical and morphological information. Their system made use of a tagger based on the finite state technology to conduct a lexical and morphological analysis of each word in the first level. The tagger tokenizes sentences, conducts a lexical lookup to process a morphological analysis and performs part-of-speech tagging. Each word in the sentence is given various tags and a special flag. The tags include noun, proper noun and abbreviation, etc. The special flag indicates whether the word matches a known word or is "guessed". Based on the tags and the special flag, they built a series of rules including recovery rules, algorithmic rules and contextual rules. The recovery rules recognize gene names using a domain-specific dictionary containing about 200 general biological expressions. The algorithmic rules recognize gene names based on prefixes (nearly 100 entries), suffixes (nearly 200 entries) and complex expressions such as nucleotide sequences and peptide notations. The contextual rules apply lexical-syntactic patterns to make final validations on candidates. Their system achieved precision of 91.4% and recall of 94.4% on a small corpus (1200 sentences) from FlyBase. However, they found that when they applied the system to a larger corpus (25,000 MEDLINE abstracts) and evaluated the performance by sampling, the precision was reduced to around 70%.

[Gaizauskas et al. 2000] derived their system from an already developed IE system in the MUC. Their system was applied in two projects: extraction of information about enzymes and metabolic pathways (EMPathIE) and extraction of information about protein structure (PASTA). Their system consisted of five processing stages: text processing, morphological analysis, term lookup, terminology parsing and term matching. The main information resources they used included case-insensitive terminology lexicons (the

component term of various categories) such as resources from public databases (SWISS-PROT, CATH and SCOP), morphological cues (standard biochemical suffixes) and hand-constructed grammar rules for each terminology class. The EMPathIE system was designed for 10 named entity classes, such as compound, element, enzyme, etc. and achieved precision of 86% and recall of 68% on 6 full journal articles. The PASTA system was designed for 13 named entity classes, such as protein, species, residue, etc. and achieved precision of 94% and recall of 88% on 52 MEDLINE abstracts.

Although these rule-based systems seem quite promising, they lack the ability of adaptation to new name classes in biomedical domain. Once a new name class is required to identify, a set of rules for the new class has to be generated manually. In fact, with the increasing number of name classes, the terms of entity names in different classes will overlap each other. Consequently, the more the number of classes is, the more difficult to construct the consistent rules. Moreover, up to now, the evaluations of these systems are only based on small corpus. [Proux et al. 1998] reported their system fails in a larger corpus. It seems that the rule-based system is not that robust and flexible.

1.3.2 Machine-Learning Approaches

Currently, machine learning-based approaches become more and more popular in biomedical name recognition. The typical works include [Nobata et al. 1999], [Collier et al. 2000], [Takeuchi and Collier 2002], [Kazama et al. 2002] and [Lee et al. 2003].

[Nobata et al. 1999] tried two classification methods and three identification methods for biomedical name recognition. The first classification method tried to induce a Naïve

Bayes classifier using conditional probabilities between word and class from the distribution of words in pre-classified domain-specific word lists. The second classification method used a decision tree approach which incorporated feature sets of part-of-speech information, character type information, and domain specific word lists. The three identification methods included shallow parsing, decision trees and statistical identification. The system tried to recognize 10 classes of biomedical names, such as protein, DNA, RNA, cell line, cell type, etc. They conducted a series of experiments by combination of each classification and identification method. The experiments showed that by using both decision tree methods for classification and identification achieved the best f-score of 56.98 to 66.24 on 100 manually annotated MEDLINE abstracts by 5-fold cross validation. The corpus was a preliminary version of GENIA corpus.

[Collier et al. 2000] applied linear interpolating HMM for gene name and gene product name recognition. They trained HMM entirely based on surface word itself and character information. The system tried to recognize 10 classes of biomedical names. The classes and the corpus were the same as those in [Nobata et al. 1999]. The system achieved f-score of 72.8.

[Takeuchi and Collier 2002] did their experiment based on SVM. The model incorporated surface word, orthographic feature and the class assignments of context words. The window size of context was -3 to +3. In their experiment, they found that part-of-speech features degraded the performance in their model. The evaluation was also conducted on the same corpus as used in [Nobata et al. 1999] and the f-score was 71.78.

[Kazama et al. 2002] developed a system also based on SVM. To our knowledge, it is one of the earliest published works on the GENIA corpus V1.1, which contains 670 MEDLINE abstracts and 24 named entity classes. Compared with [Nobata et al. 1999], they made use of richer features, such as word feature, part-of-speech feature, prefix feature, suffix feature, previous class feature, word cache feature and HMM state feature. They use a BIO (beginning/in/out of entity name) representation to classify a word. For example, a word belongs to class "B-DNA" if it is a beginning of a DNA name; a word belongs to class "I-Protein" if it is inside a protein name; and a word belongs to class "O" if it is outside a entity name. They tried to classify each word into one of such categories to represent the name recognition. Since standard SVM model is a binary classifier, the pair-wise strategy was used for constructing a multi-class SVM. In addition, they used a class splitting technique for balancing class distribution. Since there were too much samples of class "O", they combined class "O" and part-of-speech tags, and thus split class "O" into several subclasses, such as "O-NN", "O-JJ" and etc. They divided the 670 abstracts in GENIA corpus V1.1 into a training set of 590 abstracts and a test set of the rest 80 abstracts. Their system achieved f-score of 54.4.

[Lee et al. 2003] developed a two-phase system based on SVM. The main idea of their work was to separate the name recognition procedure into two phases: identification and classification. They trained an SVM model for name identification, which is to find the locations of all the names. After the identification phase, they trained another SVM model for name classification, which is to classify the identified names. The feature set for name identification consisted of part-of-speech, suffix, prefix and surface word. The feature set for name classification consisted of functional words, inside words, left context words and

right context words. They conducted experiments both on GENIA corpus V1.1 and V3.0p. However, the part-of-speech information they used in their experiments was from the annotated corpus instead of some automatic part-of-speech tagger. This could result in the unpersuasive comparisons between their results and the others'.

Certainly, it is difficult to compare various models because of the different classes and corpus they used. Considering [Nobata et al. 1999], [Collier et al. 2000] and [Takeuchi and Collier 2002] using the same class and corpus, we make a rough comparison among them. The results show that HMM and SVM outperform decision tree and the performance of HMM and SVM are almost equivalent. The results also show that how to capture the useful evidence for domain-specific names and how to integrate them effectively in the model is crucial. In this respect, [Collier et al. 2000]'s HMM model only use surface word and character information, which may not be adequate for coping with the complicated biomedical names.

1.4 Our Contribution

Our contribution to the research in biomedical name recognition can be concluded as follow.

Firstly, we provide a detailed analysis of the characteristics of biomedical names as well as the statistics-based comparisons between name recognition in the newswire domain

(MUC) and the biomedical domain. Among these, the analyses of the length distribution and the cascaded name are the author's contributions.

Secondly, we develop a rich set of features for our Hidden Markov Model-based name recognizer for the biomedical domain. We also conduct a series of experiments to evaluate the impact of different features. Especially for the part-of-speech feature, we adapt a part-of-speech tagger into the biomedical domain to get more accurate part-of-speech feature and present the necessity of such adaptation by comparisons. Among these contributions, the design of the orthographic features, morphological features and the adaptation of the part-of-speech tagger are the author's contributions.

Thirdly, our system is the first that tries to resolve the phenomena of cascaded names in the biomedical domain. In this thesis, two solutions are proposed: a post-processing rule-based approach and an HMM-based iterative recognition approach. The cascaded name recognition is the author's contribution.

In addition, we make a detailed error analysis, from which we can have certain observations for potential improvement in the future. Last but not least, we build a concrete system for biomedical name recognition. The experimental results show that our system outperforms the previous best published system by 8.1 f-measure on the same training and test data of GENIA corpus V1.1. Demo of our work can also be accessed at http://textmining.i2r.a-star.edu.sg/NLS/demo.htm. Among these, a part of the error analysis, about half of the system development and the development web demo are the author's contributions.

1.5 Organization of the Thesis

The thesis is organized as follows: Chapter 2 provides a detailed analysis of special characteristics of biomedical names. In Chapter 3, the rich feature set that we design for the biomedical domain is described in detail. In Chapter 4, we provide detailed description of our hidden Markov model-based name recognition model and present various methods for abbreviation recognition and cascaded name recognition. In Chapter 5, we show our experimental configurations and various experimental results. We also make analysis based on these results and present error analysis. Finally, in Chapter 6 we conclude this thesis with future works.

Chapter 2

CHARACTERISTICS OF BIOMEDICAL NAMES

2.1 Length

The length distribution of biomedical names is different from that of named entities in the newswire domain. We found that the variance of length of biomedical names is wider. In the biomedical documents, many names are very long, e.g. *"47 kDa sterol regulatory element binding factor"*. Most of named entities in the newswire domain are comparatively short and have less variance of length. We produced statistics on the length of biomedical names from GENIA corpus, as well as the length of entity names in newswire domain from the MUC-7 corpus. Figure 2.1 and Figure 2.2 show the length distributions of biomedical names and newswire named entities respectively.

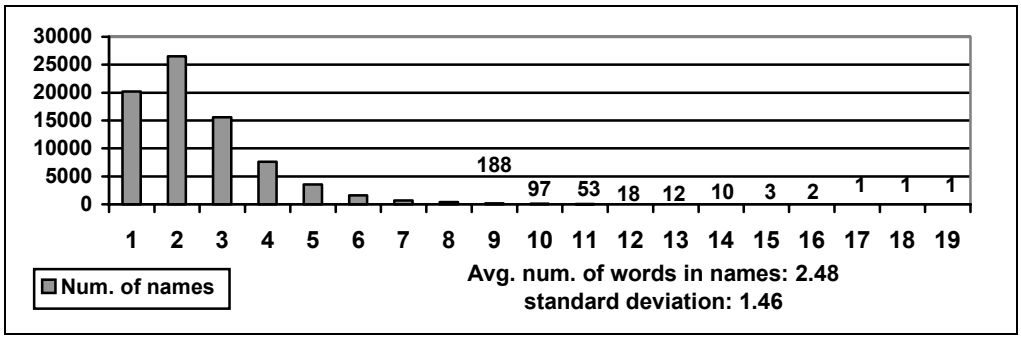


Figure 2.1: Length distribution of biomedical names in GENIA corpus V3.02

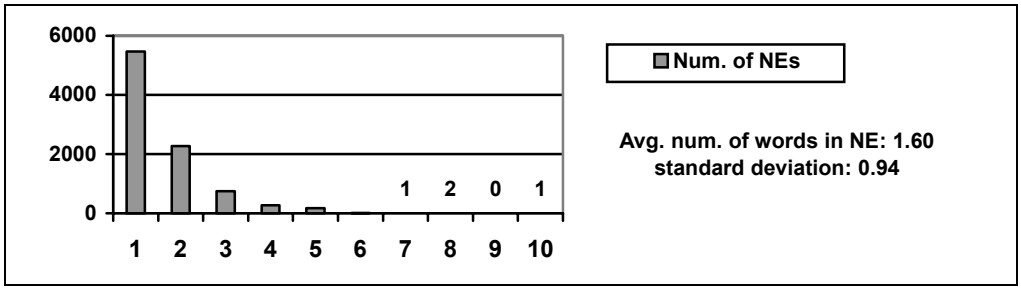


Figure 2.2: Length distribution of named entities in MUC-7 corpus

From Figure 2.1 and Figure 2.2, we can find that the average length of biomedical names is 2.48, which is greater than that of newswire named entities (1.60). In addition, newswire named entities with length over six are rare, while many biomedical names consist of more than 10 words. The standard deviation of length for biomedical names is 1.46, which is also much greater than that for newswire named entities (0.94).

2.2 Naming Conventions

One biomedical name may be found in various spelling forms, for example, *"N-acetylcysteine"*, *"N-acetyl-cyteine"*, and *"NAcetylCysteine"*. We find that the use of capitalization and hyphen is casual in biomedical documents. Naming conventions are inconsistent. On the other hand, capitalization of entity name is different between biomedical domain and newswire domain. We produced statistics on capitalization in entity names. From GENIA corpus V3.02, we find that 62.89% of words in biomedical names are in lowercase. However, from MUC-7 corpus, we find that only 1.65% of words in named entities are in lowercase. Therefore, the feature of capitalization in the biomedical domain is not as dominant as it in the newswire domain for named entity recognition.

Moreover, more and more biomedical names are constantly invented by authors and have not been collected by public databases. This may result in the inadequate coverage in some domain-specific dictionaries [Fukuda et al. 1998] [Nobata et al. 1999]. Thus, conventional named entity recognition methods which depend on pre-defined dictionaries may not perform well.

2.3 Descriptive Names

Sometimes, biomedical names are descriptive. Modifiers often occur before or after basic names to indicate property, type, nature and etc. In some cases, people disagree with each other on whether modifiers should be regarded as part of the names. That causes much inconsistency in the training data. For example, in *"normal thymic epithelial cells"*, it is hard to decide whether the word *"normal"* should be the beginning of the name.

Compared to biomedical domain, few named entities are descriptive in newswire domain. Named entities in newswire texts are mostly proper names. Even though descriptive words exist, people are only interested in the basic named entities.

2.4 Cascaded Names

Biomedical names can be compound names. One name may be embedded in another name, e.g. *"*<PROTEIN><DNA>*kappa 3*</DNA> *binding factor*</PROTEIN>*"*. Some names have more than one level of cascaded-annotation, such as "<OTHERNAME><DNA><VIRUS>*HIV-2*</VIRUS>*enhancer*</DNA> *activation*</OTHERNAME>". We produced statistics on different annotation construction, including non-cascaded-annotation, cascaded-annotation, and complicated construction (to be discussed in the next section), from GENIA corpus V3.02 (Table 2.1). From Table 2.1, we can find that 16.57% of names are cascaded-annotated and distributed over all the name classes. The ideal biomedical name recognition requires all instances of entity names be recognized,

including the embedded ones. While in practical ways, people only recognize the longest names.

In the newswire domain, only person names and location names may be embedded in other entity names, e.g. "*Eastern China Airlines*". However, as stated in the section "Nested Expressions" of the *MUC guidelines for markup of exceptional constructions*, once a person name or a location name is embedded in another name, it will be regarded as a part of that longer name and won't be annotated separately.

Annotation Construction	Percentage
Non-cascaded-annotated names	81.37%
Cascaded-annotated names	16.57%
Name with complicated constructions	2.06%

Table 2.1: Statistics on different annotation constructions from GENIA corpus V3.02

2.5 Complicated Constructions

In the biomedical domain, authors often use conjunction "and" and disjunction "or" in order to reduce redundancy. Two or more biomedical names may share one head noun. For example, *"91 and 84 kDa proteins"* consists of two names: *"91 kDa proteins" and "84 kDa proteins"*. Strictly speaking, it is required to recognize the two names separately. Some instances have more specific construction, e.g. *"IL-2- but not IL-12-stimulated NK cells"*. According to our statistics on GENIA corpus V3.02, 2.06% of biomedical names have such complicated construction, which has been shown in Table 2.1.

In the newswire domain, the section "Expressions Involving Elision" of the *MUC guidelines for markup of exceptional constructions* defined a standard for annotating complication constructions. However, these constructions have much less cases in the newswire domain.

2.6 Abbreviations

Abbreviations in the biomedical domain pose much bigger challenge to name recognition than that in the newswire domain for several reasons:

First, abbreviations in the biomedical domain are frequently used. [Chang et al. 2002] showed that, in all MEDLINE abstracts until the end of 2001, 42.8% abstracts have at least 1 abbreviation and 23.7% abstracts have two or more. They also showed that there is one new abbreviation in every 5-10 abstracts on average and the growth rate of new abbreviation is increasing. Moreover, abbreviations in the biomedical domain are well distributed in all classes, while in the newswire domain, abbreviations just occur in some classes, such as PERSON, ORGANIZATION, and rarely occur in other classes, such as TIME, PERCENT and MONEY. Therefore, the performance of the abbreviation recognition in the biomedical domain has greater impact on the overall performance.

Secondly, abbreviations in biomedical domain are formed irregularly. For instance, abbreviations may be combined from the first letters of words in the full form, e.g. *Human Immunodeficiency Virus* (*HIV*), or a subset of syllable boundaries, e.g. *Interleukin 2* (*IL2*), or several contiguous characters of a word, e.g. *palate* (*PAL*), etc. It highlights the

difficulties to build relationships between abbreviations and their full forms. While in the newswire domain, abbreviations are comparatively regular. Most abbreviations are formed by the first letters of words in the full forms, e.g. *IBM, US,* etc.

Thirdly, abbreviations in biomedical domain are highly ambiguous. For instance, *TCF* may refer to *T cell Factor* or *Tissue Culture Fluid* in different biomedical documents. [Liu et al. 2002] show that 81.2% of the abbreviations are ambiguous and have an average of 16.6 senses in MEDLINE abstracts. Name class of abbreviation depends on context and cannot be simply assigned by matching items in abbreviation dictionaries collected from public resources.

Chapter 3

FEATURES

3.1 Orthographic Features (F_o)

Orthographic features are designed to capture the word formation information, such as capitalization, digitalization and their combinations. Orthographic information have been widely used in name recognition with different design of features, such as [Zhou and Su 2002], [Nobata et al. 1999], [Gaizauskas et al. 2000], [Collier et al. 2000], [Takeuchi and collier 2002], and [Kazama et al. 2002]. Basically, orthographic features are manually designed and aim to group words by similar formation.

In the newswire domain, orthographic features are both helpful to identify locations and distinguish classes for certain names. For example, symbol '$' and '%' are good indicators of class MONEY and PERCENTAGE respectively. However, in biomedical domain, orthographic features are more likely to be served as indicators of unknown words, such as newly invented abbreviations. For example, "IL-2" is in the training data, but "IL-12" is not. Fortunately, we can guess that "IL-12" is similar to "IL-2" based on their orthographic features. Therefore, orthographic features are intuitively helpful to identify new biomedical names.

In our work, we manually designed orthographic features based on the characteristics of biomedical names. Table 3.1 shows the list of orthographic features we designed by descending order of priority.

F_o Name	Example	Explanation
Comma	,	comma
Dot	.	dot
LRB	(	left round bracket
RRB	)	right round bracket
LSB	[	left squared bracket
RSB	]	right squared bracket
RomanDigit	II, IV	Roman digit
GreekLetter	beta	Greek letter
StopWord	in, at	stop word
ATCGseq	AACAAAG	nucleotide sequence
OneDigit	5	one digit
AllDigits	60	all digits
DigitCommaDigit	1,25	digits + comma + digits
DigitDotDigit	0.5	digits + dot + digits
OneCap	T	single capital letter
AllCaps	CSF	all capital letters
CapLowAlpha	All	capital letter followed by lowercase letters
CapMixAlpha	IgM	capital letter followed by mixture of cases
LowMixAlpha	kDa	lowercase letter followed by mixture of cases
AlphaDigitAlpha	H2A	letters + digits + letters
AlphaDigit	T4	letters + digits
DigitAlphaDigit	6C2	digits + letters + digits
DigitAlpha	19D	digits + letters

Table 3.1: Sorted list of orthographic features by descending order of priority

From Table 3.1, we can find that the features such as GreekLetter, RomanDigit, ATCGseq and features dealing with mixed alphabetical letters and digits are specially designed for biomedical domain. In fact, features dealing with mixed alphabetical letters and digits such as AlphaDgtAlpha, CapMixAlpha, etc. are beneficial for biomedical abbreviations. Moreover, features such as ATCG nucleotide sequence identify the similarity of the special biomedical notations according to their orthographical forms, e.g. *AACAAAG, CTCAGGA*, etc. Besides these, some features such as comma, dot, StopWord and LSB, etc. are designed intuitively to provide information to detect the boundary of names. Especially, parentheses are often used to indicate the definition of abbreviation in

biomedical documents. Intuitively, they are useful to identify abbreviations. In Section 4.3, we will explain how to use parentheses to deal with abbreviation in detail.

3.2 Morphological Features (F_m)

Morphological information, such as prefix and suffix, is considered as an important cue for terminology identification. In our work, we use statistical method to get most frequent N prefixes and suffixes from training data as candidates. Then, each of these candidates is evaluated according to formula 4.1.

$$Wt_i = \frac{(\#IN_i - \#OUT_i)}{N_i} \qquad (4.1)$$

In formula 4.1, $\#IN_i$ is the number of times that prefix or suffix i occurs within names; $\#OUT_i$ is the number of times that prefix or suffix i occurs out of names; N_i is the total number of occurrence of prefix or suffix i.

The formula assumes that the particular prefix or suffix, which is most likely inside biomedical names and least likely outside biomedical names, may be thought as a good evidence for recognition. The candidates with Wt above a certain threshold (0.7 in experiment) are chosen. In the next step, we calculated the frequency of each prefix or suffix in each biomedical name class, and group the prefixes or suffixes with similar distributions among name classes into one feature. This is because prefixes or suffixes with similar distributions have similar contributions. The grouping procedure reduced the total number of features and prevented the model from suffering from the data sparseness problem. Table 3.2 shows some examples of the morphological features used in our work.

F_m Name	Prefix/Suffix	Example
sOOC	~cin	actinomycin
	~mide	Cycloheximide
	~zole	Sulphamethoxazole
sLPD	~lipid	Phospholipids
	~rogen	Estrogen
	~vitamin	dihydroxyvitamin
sCTP	~blast	erythroblast
	~cyte	thymocyte
	~phil	eosinophil
sPEPT	~peptide	neuropeptide
sMA	~ma	hybridoma
sVIR	~virus	cytomegalovirus

Table 3.2: Examples of morphological features

From Table 3.2, we can find that the suffixes ~cin, ~mide, ~zole have been grouped into one feature sOOC because they all have the high frequency in the biomedical name class *OTHER-ORGANIC-COMPOUND* and relatively low frequencies in the other name classes.

3.3 Part-of-Speech Features (F_{pos})

In the previous research of named entity recognition in newswire domain, part-of-speech (POS) feature was stated not useful. Part-of-speech information mainly aims to identify boundaries of names. In newswire domain, certain orthographic information such as capitalization is dominant in playing such a role, as we have already showed that nearly 98% of words in newswire named entities are capitalized. Thus, part-of-speech features may not help much and even degrade the performance in newswire domain [Zhou and Su 2002].

In biomedical domain, orthographic information is not strong enough for name identification, since many names are not capitalized. Besides this, many biomedical names are descriptive and long, identifying the name boundaries is not a trivial task. Intuitively, names are more likely to occur within noun phrases than other phrases. Part-of-speech tags can provide the evidence of the noun phrase region based on the syntactic information of the words. In our work, we will show that the part-of-speech feature is an important feature in biomedical name recognition.

However, in previous related work, [Kazama et al. 2002] made use of part-of-speech information and concluded that it could only slightly improve the performance. In [Collier et al. 2000], [Nobata et al. 2000] and [Takeuchi and Collier. 2002], part-of-speech information was not incorporated in their models. The probable reason explained by them was that the part-of-speech taggers they used were trained on the newswire documents. Thus, the part-of-speech tags may not be accurate enough in the biomedical domain. On the whole, part-of-speech information hasn't been well used in previous related work of biomedical name recognition.

To demonstrate the effect of adaptation on part-of-speech tagging, we trained an HMM-based part-of-speech tagger for biomedical domain using GENIA corpus V2.1 (670 abstracts, 123K words). We use 590 abstracts as the training set and the rest 80 abstracts as the test set. As for comparison, we also trained a part-of-speech tagger on the PENN TreeBank (2500 Wall Street Journal articles, 1M words) and evaluated on the same 80 abstracts of GENIA corpus. The results are shown in Table 3.3.

Training set	Test set	Precision
590 GENIA abstracts	80 GENIA abstracts	97.4
2500 WSJ articles		85.1

Table 3.3: Performance comparison between part-of-speech taggers using training data in different domains

From Table 3.3, we can find that the part-of-speech tagger trained in the biomedical domain performs much better on the test set than that trained on WSJ documents. This is consistent with the earlier explanation for why part-of-speech features are not so useful in biomedical name recognition [Nobata et al. 2000], [Takeuchi and Collier 2002].

In Chapter 5, we will show the effect of different part-of-speech taggers trained with different training data on the biomedical name recognition.

3.4 Semantic Trigger Features

We designed semantic trigger features in order to provide indications of certain biomedical name classes based on the semantic information of some trigger words. Trigger words are key words inside or outside of names, which have strong indication for name recognition. Initially, we collected two types of semantic triggers: head noun triggers and special verb triggers.

3.4.1 Head Noun Triggers (F_{hnt})

Head noun means the main noun or noun phrase of some compound words, which describes the function or property of these words, e.g. “B cells” is the head noun for the

name "activated human B cells". Compared with the other words in the biomedical names, head noun is a much more decisive factor for distinguishing the classes. For instances,

<OtherName>*IFN-gamma treatment*</OtherName>

<DNA>*IFN-gamma activation sequence*</DNA>

Both instances above begin with the "*IFN-gamma"* with only a difference in head nouns, "*treatment*" and "*sequence*". These two biomedical names belong to two different classes: *OTHER-NAME* and *DNA*. This example implies that no matter how many similar expressions are within the names, the classes of the names are normally determined by the head nouns which often indicate the functions of the names. [Nobata et al. 1999] also argued that head nouns in noun phrases can provide significant clues about the class and gave head noun a higher weight in their statistical model.

Name class	Unigram	Bi-grams head noun
PROTEIN	kinase interleukin interferon ligand …	binding protein activator protein gene product cell receptor …
CELL TYPE	lymphocyte astrocyte fibroblast eosinophil …	blast cell blood lymphocyte peripheral monocyte killer cell …
DNA	DNA cDNA chromosome breakpoint …	X chromosome alpha promoter binding motif promoter element …
VIRUS	virus provirus cytomegalovirus adenovirus …	recombinant virus lymphotropic herpesvirus virus particles immunodeficiency virus …

Table 3.4: Examples of auto-generated unigram and bi-grams head noun triggers for 4 classes

In our work, head noun features were retrieved automatically, which made our model easy to be adapted to a new domain. We extracted unigram and bi-grams of head nouns automatically from training data and rank them by frequency. For each name class, we selected 60% top ranked head nouns for both unigram and bi-grams head noun trigger lists. A sample list of auto-generated head nouns for certain name classes is shown in Table 3.4. In the future work, head nouns may also be extracted from some public resources for further enhancement.

3.4.2 Special Verb Triggers (F_{svt})

Besides collecting the trigger words inside biomedical names, such as head noun triggers, we can also make use of trigger words from the local context of names. Recently, some frequently occurred verbs in MEDLINE have been proven useful for extracting the interaction between biomedical entities, e.g. the protein-protein interactions [Thomas et al. 2000], [Sekimizu et al. 1998]. In biomedical name recognition, we have the intuition that particular verbs may also provide information for boundary and class of biomedical names. For instance, the verb "*bind*" often indicates interaction between proteins.

In our work, we selected 20 most frequent verbs which occur adjacent to biomedical names from training data automatically as the special verb trigger features, which are shown in Table 3.5.

Special Verb Trigger				
activate	associate	bind	block	clone
demonstrate	express	identify	increase	induce
inhibit	investigate	involve	isolate	mediate
observe	reduce	regulate	reveal	stimulate

Table 3.5: 20 top frequent special verb triggers

Chapter 4

METHODOLOGY

4.1 Hidden Markov Model

Hidden Markov Model (HMM) is a statistical method. In the past fifteen year, HMM has been successfully used in a wide range of applications, such as speech recognition and natural language processing. A HMM is a model where a sequence of outputs is generated in addition to the Markov state sequence. It is a latent variable model in the sense that only the output sequence is observed while the state sequence remains "hidden".

In name entity recognition, the input word sequence, (e.g. sentence), can be regarded as the observation sequence and the output tag sequence is the optimal state sequence corresponding to the words.

4.2 Hidden Markov Model-based Name Recognizer

In our work, the name recognizer is adapted from the previous work of the HMM-based Named Entity Recognizer on MUC [Zhou and Su 2002]. The core technique is a Hidden Markov Model as follows.

The name recognizer tries to find the most likely tag sequence $T_1^n = t_1 t_2 \cdots t_n$ for a given sequence of tokens $O_1^n = o_1 o_2 \cdots o_n$ that maximizes $P(T_1^n \mid O_1^n)$. In token sequence O_1^n, the

token o_i is defined as $o_i =< f_i, w_i >$, where w_i is the word and f_i is the feature sets corresponding to the word w_i. The feature sets have been introduced in Chapter 3 in detail. In tag sequence T_1^n, each tag t_i is structural and consists of three parts: boundary category, name class and feature set. The boundary category indicates whether the word itself is a name, or the word is at the beginning, in the middle, or at the end of a name. The name class consists of a NOT-NAME class and a predefined set of name classes. The feature set is added in order to represent more accurate models. [Zhou and Su 2002]

In the model, $P(T_1^n \mid O_1^n)$ can be represented as:

$$\log P(T_1^n \mid O_1^n) = \log P(T_1^n) + \log \frac{P(T_1^n, O_1^n)}{P(T_1^n) \cdot P(O_1^n)} \qquad (4.1)$$

The second term of the right-hand side of equation (4.1) is the mutual information between T_1^n and O_1^n. We assume mutual information independence:

$$\log \frac{P(T_1^n, O_1^n)}{P(T_1^n) \cdot P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(t_i, O_1^n)}{P(t_i) \cdot P(O_1^n)} \qquad (4.2)$$

Applying equation (4.2) to equation (4.1), we have:

$$\log P(T_1^n \mid O_1^n) = \log P(T_1^n) - \sum_{i=1}^{n} \log P(t_i) + \sum_{i=1}^{n} \log P(t_i \mid O_1^n) \qquad (4.3)$$

The first term in equation (4.3) can be computed by applying chain rules. Each tag is assumed to be probabilistically dependent on the N-1 previous tags in N-gram modeling. The second term is the sum of log probabilities of all the tag instances. The third term can be ideally estimated by the forward-backward algorithm recursively [Rabiner 1998]. As we designed a rich feature set in HMM, we will encounter the data sparseness problem.

An alternative back-off modeling approach by means of constraint relaxation was applied in our model [Zhou and Su 2002]. It enables the decoding process effectively find a near optimal frequently occurred pattern entry in determining the tag probability distribution of current word.

The Viterbi algorithm [Viterbi 1967] is implemented to find the most likely tag sequence in the state space of the possible tag distribution based on the state transition probabilities. Meanwhile, some constraints on the boundary category and entity category between two consecutive tags are applied to filter the invalid name tags.

4.3 Abbreviation Recognition

In Chapter 2, we have analyzed the characteristics of abbreviations in the biomedical domain. We find that it is difficult to recognize biomedical abbreviations in documents. Therefore, specific methods should be developed for recognizing them. Since most abbreviations have special orthographic formations, it is relatively easy to identify their locations. However, it is difficult to determine whether they are biomedical abbreviations or other abbreviations. If they are biomedical abbreviations, it is still a real challenge to determine which name classes they belong to.

In our work, we introduce a method to recognize abbreviations by mapping them to their full forms. This approach is based on the assumption that full forms are easier to be classified than abbreviations. In most cases, this assumption is valid because the full forms have more features than their abbreviations for classification. Therefore, if the

mapping is successful, the classes of abbreviations can be determined from the classes of their full forms, which are more likely to be correct. In a document-level point of view, if we can map the abbreviations to their full forms in the current document, the recognized abbreviations will be helpful for classifying the same forthcoming abbreviation throughout the document using the name alias feature, as in [Zhou and Su 2002].

In biomedical documents, abbreviations are often defined first before they are used. [Schwartz and Hearst 2002] found that abbreviations and their full forms occur together with a pair of parentheses in most cases. Normally, there are two patterns:

1. full form (abbreviation)
2. abbreviation (full form)

Most occurrences of abbreviations conform to the first pattern. If there are more than two words in the parenthesis, the second pattern is assumed [Schwartz and Hearst 2002].

In our name recognition model, we specially treat parentheses since they can be both useful and confusing. On one hand, parentheses strongly indicate location of abbreviations and help to map abbreviations to their full forms. On the other hand, parentheses will sometimes confuse the recognition model because they make the construction of names more complex. Normally, an abbreviation follows its full form, such as "*<Protein>T cell factor</Protein>* ***(<Protein>TCF</Protein>)***". However, since biomedical names can be cascaded, an abbreviation may be embedded in a longer name, for example, *"<DNA>chloramphenicol acetyltransferase* ***(CAT)*** *gene</DNA>"*.

Based on the characteristics of the parentheses, we design our biomedical abbreviation recognition method which takes advantage of the usefulness as well as reduces the confusion of parentheses. We reduce the complexity of sentence by extracting out all parentheses pairs in the first step. Therefore, parentheses will not affect the name recognition in that sentence. After the recognition, all parentheses items are restored back to the sentence. Now, we can take advantage of the usefulness of the parentheses to recognize abbreviations by analyzing the relationship between full forms and abbreviations. The algorithm of our biomedical abbreviation recognition method is presented in Figure 4.1.

For each sentence S_i in the document{
  if exist parenthesis{
    judge the case of {
    "full form (abbr.)";
    "abbr. (full form)";
    }
    store the abbr. A and position P_a to a list;
    record the parenthesis position $P_{p;}$
    remove A and parenthesis from sentence;
    apply HMM-based name recognizer to S_i;
    restore A and parenthesis into P_a, P_p;
    if P_p within an recognized name E with the class C_E
      parenthesis is included in E;
    else{
      parenthesis is not included;
      classify A to C_E;
      classify A in the rest part of document to C_E;
    }
  }
  else apply HMM-based name recognizer to S_i;
}

Figure 4.1: Biomedical abbreviation recognition algorithm

4.4 Cascaded Name Recognition

As mentioned in Section 2.5, cascaded-annotation is a special problem in biomedical name recognition. For instance, *"*<CELL-LINE><VIRUS>*HTLV-I*</VIRUS>*-infected cord blood lymphocytes*</CELL-LINE>*"* is a name of class *CELL-LINE* and it embeds a virus name *"HTLV-I"*. In cascaded name recognition, we shall recognize both the embedded name and the longest name. However, in real applications, people care more about names to the longest extent for two reasons. First, names to the longest extent are more likely to be the subjects that people want to study. Second, it keeps all information about embedded names. Therefore, whether keeping the embedded names or not depends on the user requirements under different circumstances. In our work, we tried two approaches: post-processing rule-based approach and HMM-based approach.

4.4.1 Post-Processing Rule-based Approach

The post-processing rule-based approach aims to recognize names to the longest extent. For a one-pass recognition procedure, it is not guaranteed that the longest names will be recognized. Instead, embedded names are sometimes recognized. For example:

… cocultured with a ***bone marrow****-derived stromal cell line revealed …*
tissue *cell_line*

In the above case, the embedded tissue name *"bone marrow"* is recognized, while we prefer to recognize the name of the cell line. The possible reason for this kind of cases is that embedded name can be individual name itself, and it contains features for itself to be recognized. The word *"marrow"* is a head noun for name class *TISSUE*, thus *"bone marrow"* is likely to be recognized as a tissue name. Although *"cell line"* is the head noun for name class *CELL-LINE*, the context is too long to be captured.

We proposed a post-processing rules approach to deal with these cases caused by cascaded-annotation. The main idea is that we try to develop a set of patterns which help recognize names to the longest extent based on embedded ones. From GENIA corpus annotation, we collected four basic patterns of cascaded names (Table 4.1). In addition, we also extend the patterns by combining the basic ones iteratively.

Basic patterns
<NAME'> = <NAME> [head nouns] <NAME'> = [modifier] <NAME> <NAME'> = <NAME^1> <NAME^2> <NAME'> = <NAME^1> [words] <NAME^2>
Extended patterns
<NAME'> = [modifier] <NAME> [head nouns] <NAME'> = [modifier] <NAME^1> <NAME^2> <NAME'> = <NAME^1> <NAME^2> [head nouns] …
Table 4.1: Patterns of cascaded names

Based on these patterns, we can construct a rule set automatically from the training corpus. Table 4.2 shows an example list of instances of the post-processing rules which were generated from the training data of GENIA corpus.

Rule instance	<DNA> = <PROTEIN> binding site
From pattern	<NAME'> = <NAME> [head nouns]
Example	*A **Myc-associated zinc finger protein binding site** is one of …*
Rule instance	<PROTEIN> = <VIRUS> <PROTEIN>
From pattern	<NAME'> = <NAME^1> <NAME^2>
Example	*Nevertheless, the **simian EBV LMP1s** retain most functions in …*
Rule instance	<CELL-TYPE> = human <CELL-TYPE>
From pattern	<NAME'> = [modifier] <NAME>
Example	*… suggests that **human NK cells** provide an effective …*
Rule instance	<CELL-TYPE> = <VIRUS> -infected <CELL-TYPE>
From pattern	<NAME'> = <NAME^1> [words] <NAME^2>
Example	*… p24 production by **HIV –infected human macrophages** when …*

Table 4.2: Examples of post-processing rules for cascaded name recognition

After name recognition by our HMM-based recognizer, we get an initial result. Then, we develop a post-processing procedure by applying the above rules to the initial result. For example, given the initial result *"a* <PROTEIN>*Myc-associated zinc finger protein*</PROTEIN> *binding site is one of ..."*, the post-processing procedure finds that it matches the rule "<DNA> = <PROTEIN> binding site". After post-processing, the final result will change to *"a* <DNA>*Myc-associated zinc finger protein binding site*</DNA> *is one of ..."*. Therefore, we recognized the name to the longest extent. In addition, the post-processing rules are applied iteratively until no new match can be found. For example, given the initial result "... <AMINO-ACID-MONOMER>*tyrosine*</AMINO-ACID-MONOMER> *kinase inhibitor* ...", the post-processing procedure finds that it matches the rule "<PROTEIN> = <AMINO-ACID-MONOMER> kinase" and changes it to "... <PROTEIN>*tyrosine kinase*</PROTEIN> *inhibitor* ..." in the first iteration. In the next iteration, the post-processing procedure finds that the intermediate result matches the rule "<OTHER-ORGANIC-COMPOUND> = <PROTEIN> inhibitor" and updates it again. Since no more matches will occur in the following iterations, the final result is "... <OTHER-ORGANIC-COMPOUND>*tyrosine kinase inhibitor*</OTHER-ORGANIC-COMPOUND > ...".

Experimental result showed that post-processing rules approach is an effective approach. Detailed results will be shown in Chapter 5.

4.4.2 HMM-based Iterative Recognition Approach

Besides post-processing rules approach, we also proposed an HMM-based iterative recognition approach to recognize names with cascaded-annotation. The HMM-based

approach tries to start recognition from the shortest embedded name and extends to longer name iteratively.

We train two HMM models in this approach. First model is the model of our original name recognizer, which is mainly for the recognition of the short embedded names. Besides this, we also train another HMM model for iterative extension for longer names by using cascaded-annotation in the training corpus. For training this iterative model, we make use of the cascaded-annotations and transform them into a new training data set. For example:

<DNA>**<PROTEIN>*Myc-associated zinc finger protein***<PROTEIN>** *binding site*</DNA>

<DNA> **[PROTEIN]** *binding site*</DNA>

We substitute a class-representing token "[PROTEIN]" for the embedded name of protein. After this transformation, all cascaded-annotated names in the training data become non-cascaded. We can train an HMM model on this training data as an iterative recognition model. Intuitively, the HMM model can capture local context information more easily than long context information. Some long cascaded names are difficult to be recognized in one pass as shown in the previous section. We hope that long cascaded names can be recognized by two or more iterations if they are missed in the first pass. Therefore, we can use the same HMM method and do not need any post-processing step. One limitation of the approach may be that the following iterations rely on the first recognition pass. In an ideal situation, if the performance is high in the first pass, the longer names are likely to be recognized in the following iterations. In our work, we are concerning about the performance of the longest names, the performance evaluation in the experiment is

conducted on them. The algorithm of the HMM-based iterative recognition approach is shown in Figure 4.2.

```
for each sentence Si in the document{
    apply the first pass HMM name recognition model to Si ;
    for each recognized name Nj {
        record Nj to a stored-list;
        substitute a class-label token CT(Nj) for Nj in Si ;
    }
    loop until no name can be recognized in Si{
        apply the iterative recognition HMM model to Si ;
        for each Nj in the stored-list {
            if CT(Ni) is embedded in newly recognized name N'k {
                restore content of Nj to original position CT(Nj) in Si ;
                remove Nj from the stored-list;
            }
        }
        for each newly recognized name N'k {
            record N'k to a stored-list;
            substitute a class-label token CT(N'k) for N'k in Si ;
        }
    }
}
```

Figure 4.2: Algorithm of HMM-based iterative recognition approach for cascaded name recognition

In addition, we can also generalize the model to a recursive process which can be used to recognize all levels of the cascaded-names, i.e. not only the longest names but also the embedded ones. The algorithm for this generalized method is shown in Figure 4.3.

for each sentence S_i in the document{
    apply first pass HMM name recognition model to S_i
    for each recognized name N_j {
        record N_j to a stored-list;
        substitute a class-label token $CT(N_j)$ for N_j in S_i
    }
    recursive-recognize-cascaded-name(S_i);
    for each N_j in the stored-list {
        restore N_j to original position $CT(N_j)$ in S_i ;
    }
}

function recursive-recognize-cascaded-name(sentence S){
    apply the iterative recognition HMM model to S ;
    if no name is recognized **then return** ;
    for each recognized name N_i {
        record N_i to a local stored-list;
        substitute a class-label token $CT(N_i)$ for N_i in S ;
    }
    recursive-recognize-cascaded-name(S);
    for each N_i in the stored-list {
        restore N_i to original position $CT(N_i)$ in S ;
    }
}

Figure 4.3: Algorithm of a generalized recursive method for all-level cascaded name recognition

4.5 Work Flow of Biomedical Name Recognition

In the whole process of biomedical name recognition, the core HMM-based name recognizer module is collaborated with other modules, including a preprocess module, a part-of-speech tagger, and a cascaded names recognition module. The work flow of the process is characterized in Figure 4.4.

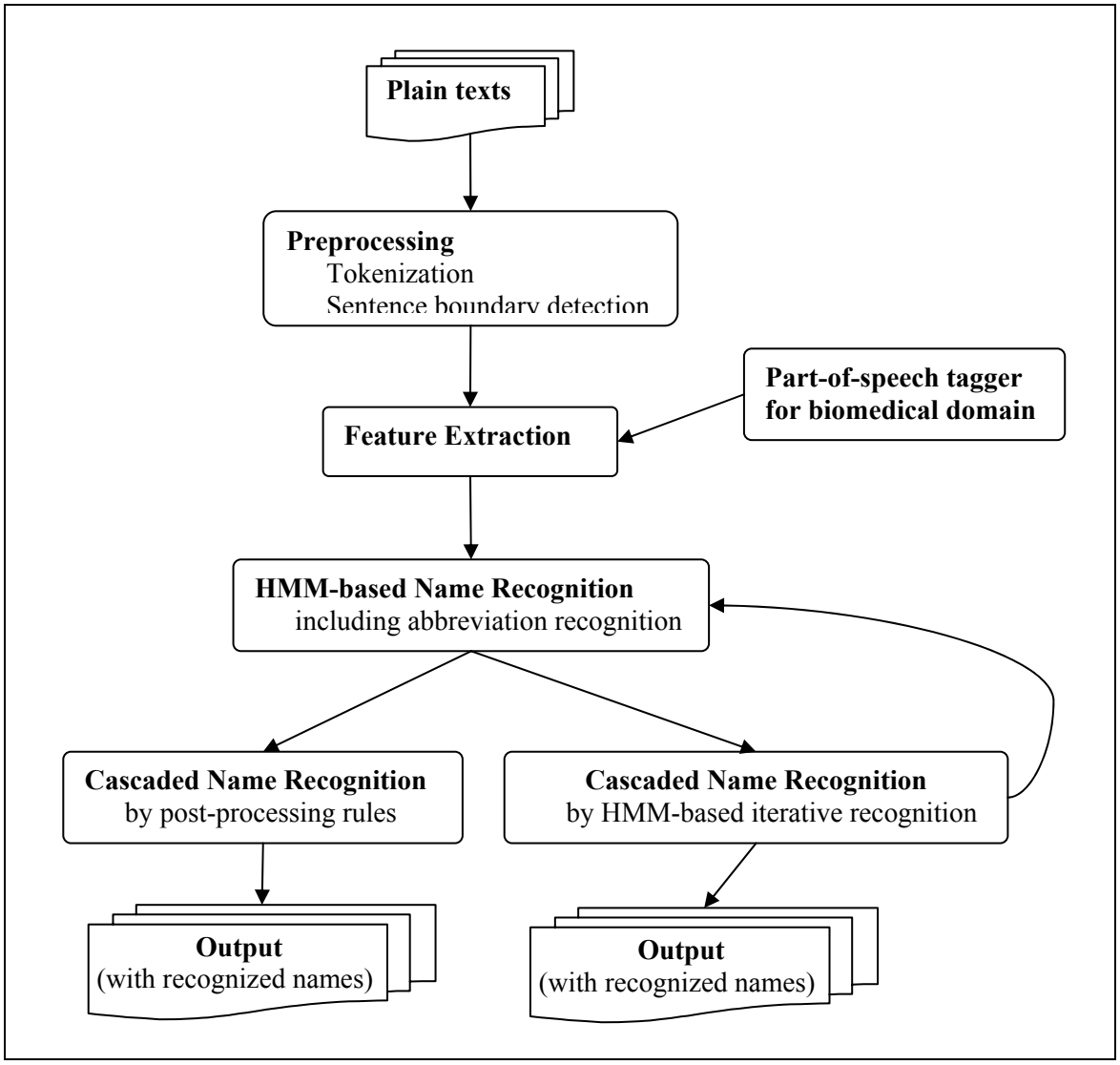


Figure 4.4: Flowchart of biomedical name recognition

Chapter 5

EXPERIMENTATION

5.1 Data set

Currently, GENIA corpus[1] is the largest annotated corpus in molecular biology domain available to public [Ohta et al. 2002]. In our experiment, three versions are used, which are V1.1, V2.1 and V3.02. The annotation of biomedical name is based on the GENIA ontology [Appendix]. In our task, we try to recognize 22 distinct name classes[2] defined in the GENIA ontology, including *MULTI-CELL, MONO-CELL, VIRUS, BODY-PART, TISSUE, CELL-TYPE, CELL-COMPONENT, CELL-LINE, OTHER-ARTIFICIAL-SOURCE, PROTEIN, PEPTIDE, AMINO-ACID-MONOMER, DNA, RNA, POLYNUCLETOTIDE, NUCLEOTIDE, LIPID, CARBOHYDRATE, OTHER-ORGANIC-COMPOUND, INORGANIC, ATOM* and *OTHER*.

5.1.1 GENIA corpus V1.1

It contains 670 MEDLINE abstracts. Since a lot of previous related works were based on this version, we use it to compare our result with others' works.

5.1.2 GENIA corpus V2.1

It contains the same 670 abstracts as V1.1 with additional part-of-speech tagging. We use this version of corpus to adapt the part-of-speech tagger to the biomedical domain and make evaluations.

[1] Downloaded from http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[2] In previous work on GENIA V1.1, there are 23 name classes due to inconsistent annotations of class *ORGANISM*. According to GENIA ontology, *ORGANISM* is not a name class in V3.02. We do not differentiate the subclasses of *PROTEIN, DNA* and *RNA*.

5.1.3 GENIA corpus V3.02

It contains 2000 MEDLINE abstracts, which is a superset of GENIA corpus V1.1. We use this version of corpus to evaluate our system more consistently and to find out the effect of training data size.

5.2 Experiments for Biomedical Name Recognition

5.2.1 Experiment Settings

We conducted experiments for biomedical name recognition both on GENIA corpus V1.1 and V3.02. For GENIA corpus V1.1, we split the corpus into a training set of 590 abstracts and a test set of 80 abstracts. We keep the same training/test ratio as [Kazama et al. 2002] in order to make comparisons. For GENIA corpus V3.02, 2000 abstracts were split to a training set of 1920 abstracts and a test set of 80 abstracts. The test set in this setting is the same as the test set for GENIA corpus V1.1 in order to evaluate the impact of training size. As a summary, the settings for biomedical name recognition are shown in Table 5.1.

	V1.1	V3.02
Training set	590 abstracts	1920 abstracts
Test set	80 abstracts	

Table 5.1: Experiment settings for biomedical name recognition on GENIA corpus V1.1 and V3.02

5.2.2 Experiment Results

The performance of our model is evaluated using "precision/recall/f-measure", in which "precision" is calculated as the ratio of the number of correctly found names to the total number of names found by our model; "recall" is calculated as the ratio of the number of

correctly found names to the number of true names; and "f-measure" is defined by Formula 5.1.

$$f-measure = \frac{2 \times precision \times recall}{precision + recall} \quad (5.1)$$

The experiment results on the overall performance of the three settings and the result of [Kazama et al. 2002] are shown respectively in Table 5.2.

	Precision	Recall	F-measure
Our model on V3.02	67.7	65.3	66.5
Our model on V1.1	63.8	61.3	62.5
[Kazama et al. 2002] on V1.1	56.2	52.8	54.4

Table 5.2: Overall performance of biomedical name recognition on GENIA corpus V3.02 and V1.1, comparing to [Kazama et al. 2002] on GENIA corpus V1.1.

On V1.1, Our system (62.5 F-measure) outperforms [Kazama et al. 2002] (54.4 F-measure) by 8.1 F-measure. It probably benefits from the various evidential features and the effective methods we proposed. Furthermore, as expected, the performance achieved (66.5 F-measure) on V3.02 is better than that on V1.1 (62.5 F-measure), which indicate that training data size matters. In the later section 5.2.5, we will show a detailed experiment result on the effect of training data size.

Besides the overall performance, we also evaluated performances of all the name classes, which are shown in Table 5.3. From Table 5.3, we can find that performance varies a lot among different name classes. It is probably due to two reasons. First, different name classes have different difficulties for name recognition. For example, the name class *BODY-PART* is one of the easiest classes since the number of instances for body part is limited. Second, we can find that the numbers of training and test instances are not evenly

distributed among all the name classes. Some minor name classes, such as *NUCLEOTIDE, ATOM, INORGANIC, CARBOHYDRATE* and etc, lack enough training or test data to achieve acceptable performances.

Name class	# Train Instance	# Test Instance	P	R	F
BODY-PART	378	16	73.68	87.50	80.00
CELL-TYPE	6219	212	78.18	81.13	79.63
LIPID	1617	78	74.68	75.64	75.16
MULTICELL	1516	99	68.47	76.77	72.38
PROTEIN	24493	790	71.88	72.15	72.01
OTHER-ORGANIC-COMPOUND	3487	106	74.00	69.81	71.84
CELL-LINE	3658	134	68.85	62.69	65.62
AMINO-ACID-MONOMER	361	25	64.00	66.66	65.30
DNA	8376	474	66.90	60.13	63.33
CELL-COMPONENT	602	33	59.46	66.67	62.86
POLYNUCLEOTIDE	234	15	58.82	66.67	62.50
OTHER-NAME	19131	702	62.17	61.11	61.64
RNA	698	69	82.50	47.83	60.55
VIRUS	1009	63	75.61	50.17	60.32
TISSUE	625	15	66.67	53.33	59.26
MONO-CELL	177	2	50.00	50.00	50.00
NUCLEOTIDE	126	6	100.00	16.67	28.57
PEPTIDE	398	9	14.29	11.11	12.50
ATOM	156	3	0	0	0
INORGANIC	210	1	0	0	0
CARBOHYDRATE	69	0	0	0	0
OTHER-ARTIFICIAL-SOURCE	199	0	0	0	0

Table 5.3: Results for biomedical name recognition of each name class on GENIA corpus V3.02.

5.2.3 Effectiveness of Feature Sets

In order to evaluate the contribution of each feature set, we conducted experiments on various combinations of features on V3.02. The results are shown in Table 5.4.

From Table 5.4, we analyze the contribution of each feature in the following part:

1) Based on orthographic feature (F_o), our system achieves a basic level performance of 29.4 F-measure. In MUC-7 task, performance can reach 77.6 F-measure by using orthographic feature only [Zhou and Su 2002]. It suggests that in biomedical domain orthographic feature is not so informative, which is consistent with what we analyzed in the Chapter 3.

F_o	F_m	F_{pos}	F_{hnt}	F_{svt}	Precision	Recall	F-measure
√					41.8	21.8	28.7
√	√				44.4	24.3	31.4
√		√			55.7	49.4	52.3
√			√		55.9	44.9	49.8
√	√	√			58.0	51.3	54.5
√	√		√		55.8	44.8	49.7
√		√	√		61.9	61.5	61.7
√	√	√	√		**61.9**	**61.7**	**61.8**
√	√	√	√	√	60.6	59.3	60.0

Table 5.4: Experimental results for biomedical name entity recognition by using different combinations of features.

2) Morphological feature (F_m) led to positive effect by +2.7 F-measure improvement based on F_o and +2.2 F-measure improvement based on F_o+F_{pos}. However, it cannot make improvement based on F_o+F_{hnt} and can only slightly improve the Recall by +0.2 based on F_o+F_{pos}+ F_{hnt}. The probable reason is that F_m and F_{hnt} provide some overlapping information. The evidences included in F_m probably can also be captured by F_{hnt}. Moreover, the evidences captured by F_{hnt} are more accurate than that captured by F_m. The contribution made by F_m may come from where there is no indication of F_{hnt}.

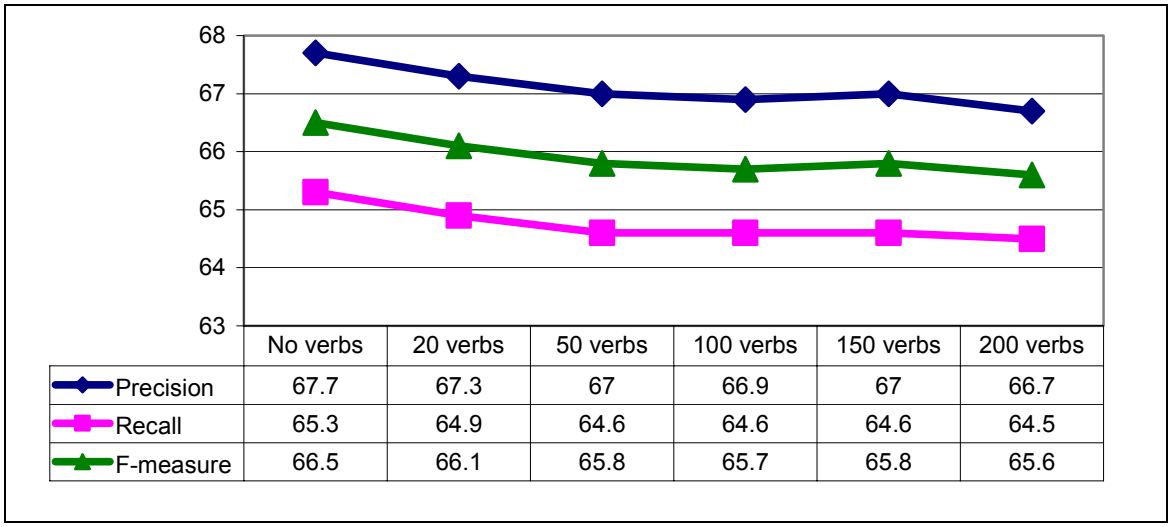
3) Part-of-speech feature (F_{pos}) is proved very beneficial as it makes significant improvement on F-measure (+23.6 based on F_o; +23.1 based on F_o+F_m; +11.9 based on F_o+F_{hnt}; +12.1 based on F_o+F_m+F_{hnt}). This is greatly benefited from the adaptation of the

part-of-speech tagger to the biomedical domain. In order to show the effect of the adaptation of part-of-speech, we made further experiments by using part-of-speech features assigned by part-of-taggers trained on different corpus. Table 5.5 shows the results of the experiments.

	Precision of part-of-speech tagger					
	No POS	71.31	83.49	85.10	97.17	97.40
Precision	62.2	64.2	65.6	66.3	66.9	**67.7**
Recall	48.2	55.7	62.7	63.5	64.1	**65.3**
F-meaure	54.3	59.6	64.1	64.9	65.5	**66.5**

Table 5.5: Effect of adaptation of POS tagger on biomedical name recognition
(The POS tagger with the precision 85.10 is trained on 2500 WSJ articles; The POS tagger with precision 97.40 is trained on GENIA corpus V2.1; The POS tagger with precision 97.17 is trained on a combined training set of WSJ articles and GENIA corpus V2.1; The other POS taggers are trained on subsets of 2500 WSJ articles. The biomedical name recognition is based on Fo+Fm+Fpos+Fhnt and abbreviation recognition and post-processing-rule cascaded name recognition method)

4) Head noun trigger feature (F_{hnt}) has also been proved very useful. It greatly improves the F-measure (+21.1 based on F_o; +18.3 based on F_o+F_m; +9.4 based on F_o+F_{pos}; +7.3 based on $F_o+F_m+F_{pos}$).



	No verbs	20 verbs	50 verbs	100 verbs	150 verbs	200 verbs
Precision	67.7	67.3	67	66.9	67	66.7
Recall	65.3	64.9	64.6	64.6	64.6	64.5
F-measure	66.5	66.1	65.8	65.7	65.8	65.6

Figure 5.1: Effect of using special verb triggers on biomedical name recognition

5) Out of our expectation, the special verb trigger feature (F_{svt}) decreases both precision and recall and degrades the F-measure (-1.8 based on F_o+ F_m+F_{pos}+ F_{hnt}). Since we just used 20 special verbs and did not consider abbreviation and cascaded recognition methods in this evaluation, we conducted further experiments using all features, abbreviation recognition method and post-processing rule approach for cascaded name recognition to show the effect of special verb trigger size on the whole system. Figure 5.1 shows the result table and chart of the performances by using different number of special verbs as F_{svt}.

From the Figure 5.1, we find that the more special verbs we use, the worse the performance is. It indicates that special verb triggers are not helpful in our model. One possible reason is that the present and past participles of some special verbs often play the adjective-like roles inside biomedical names, such as "*IL10-inhibited lymphocytes*". Another possible reason is that the function of special verbs is very unpredictable in the corpus even the verbs are very informative in the context. For example:

<CELLCOMPONENT>*B95-8 cytosol*</CELLCOMPONENT> ***inhibited*** *specific binding ...*
... nuclear activity in <CELLLINE>*IL-10* ***-inhibited*** *lymphocytes*</CELLLINE> *…*
... and TLCK ***inhibited*** <OTHERNAME>*LPS induction*</OTHERNAME> *of ...*

5.2.4 Effectiveness of Methods for Abbreviation and Cascaded Names

In order to evaluate our proposed methods for abbreviation and cascaded name recognition, we made further experiments based on the four feature sets which led to the best performance as shown in the previous section. The results are summarized in Table 5.6.

V3.02	Precision	Recall	F-measure
$F_o+F_m+F_{pos}+F_{hnt}$ (4F)	61.9	61.7	61.8
4F+abbreviation recognition	63.4	62.7	63.0
4F+abbr.+post-processing rule-based app.	**67.7**	**65.3**	**66.5**
4F+abbr.+ HMM-based iterative recognition app.	65.5	63.0	64.2

Table 5.6: Effectiveness of abbreviation recognition method and two cascaded name recognition methods

First, we evaluated the contribution of abbreviation recognition method. The result showed that the method led to an improvement on F-measure by 1.2 based on the best combination of features $F_o+F_m+F_{pos}+F_{hnt}$ (4F). The reason why the improvement was not so significant is that our abbreviation recognition method mainly relies on the recognition of its full form. Once the full form is wrongly recognized, all abbreviations can be wrong altogether. However, the principle of the method is reasonable and the result is positive. Our abbreviation recognition method provides an effective and reasonable solution when domain-specific abbreviation dictionaries are not available.

Furthermore, we evaluated the two approaches for recognition of cascaded names that we proposed in section 4.4. Using post-processing rule-based approach, we got a significant improvement by 3.5 F-measure. Another approach, the HMM-based iterative recognition approach, also led to a positive effect of +1.2 F-measure. We can find that the post-processing rule-based approach outperforms the HMM-based iterative recognition approach. It is probably because the HMM-based iterative recognition approach does not have enough training data on the cascaded name phenomena to get a reliable performance. However, the HMM-based iterative recognition approach is more general and can be enhanced when we have enough training instances for the cascaded name phenomena.

5.2.5 Effect of Training Data Size

One important issue is about the effect of different training data size. Figure 5.2 shows the learning curve of our name recognition model with different training data size. From Figure 5.2, we can find our system still has some room for improvement with the larger training data set. This is probably because our HMM model can better capture local context dependence and better resolve the sparseness problem. However, although it is always true that more training data may lead to better performance, the more training data than about 200K words will not help much. Figure 5.2 shows nearly 200K words of training data achieve the performance of 61.4 F-measure while adding to 450K words slightly increases the performance. Therefore, in order to achieve reliable performance, about 200K words of training data may be required.

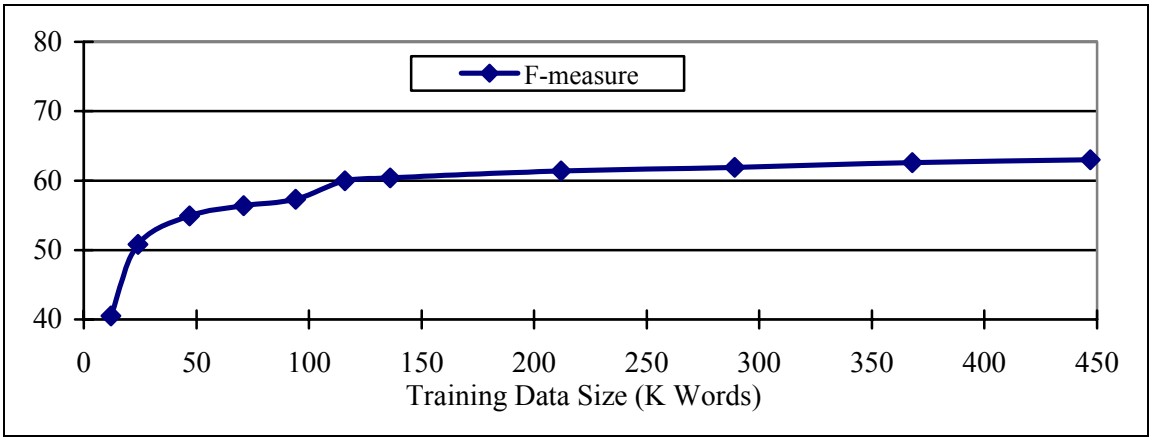


Figure 5.2: Effect of training data size for biomedical name recognition

5.3 Error Analysis

We randomly collected 100 error instances from the result of biomedical name recognition and produced a statistics on different types of errors. According to our analysis, we sort the error instances into six types, including false positives, false negatives (miss),

misclassifications, modifier-caused errors, cascaded-annotation-caused errors and none-of-the-above. Table 5.7 shows the statistics table of the error samples. In the following section, each type of error will be analyzed in details.

Error Type	# of instance
False Positive	17
False Negative (Miss)	33
Misclassification	18
Modifier-caused Error	14
Cascaded-annotation-caused Error	14
Miscellaneous Error	4
TOTAL	100

Table 5.7: Statistics of error types for biomedical name recognition by sampling 100 error instances

5.3.1 False Positive

False positive includes all wrongly recognized names that are not in the annotated evaluation data. There are 17 instances out of the 100 samples. This type of errors happens mostly in the name class *OTHER-NAME* (10/17), which is probably because there are many inconsistent annotations in the GENIA corpus and inconsistency is the most serious in name class *OTHER-NAME*. However, among all the 17 error instances, we find that 11 instances are quite acceptable. We say these instances are acceptable because we can find similar annotations in the corpus, which means inconsistency exists in annotations of the GENIA corpus. For example, in one sentence, our model recognized a false positive instance of the name *"lymphocyte activation"* as a name of class *OTHER-NAME*, while we can find that there are many *"lymphocyte activation"* annotated as a name of class *OTHER-NAME* in other places of the corpus. Therefore, this error is due to annotation error and should be regarded as acceptable.

5.3.2 False Negative

False negative errors account for the instances missed by our recognition model. From the statistics, we find one third of the errors belong to false negative errors (33/100). We find that the names missed by our model are relatively short and some of them are general nouns. For example, we missed the name "*expression*" in "*determining the patterns of* <OTHERNAME>*expression*</OTHERNAME>. *"Expression"* is a very general word in text. Most of the single word "*expression*" is not annotated as a name itself. We regard such false negatives as acceptable instances. We find that 10 out of the 33 instances are acceptable.

5.3.3 Misclassification

Misclassification accounts for right boundary detection but wrong class assignment. We find that names in some classes are sense ambiguous. In the corpus, one name can belong to two classes. From the statistics, 18 error instances are misclassification errors. Among them, 6 instances are related to the class *DNA* and 6 instances are related to the class *CELL-LINE* and the class *CELL-TYPE*. We find only 2 instances are acceptable due to wrong annotation in the corpus. All other instances are unacceptable.

5.3.4 Modifier-caused Error

Modifier-caused error accounts for correct right boundary detection, correct class assignment, but wrong identification on modifier. This type error is mainly introduced by the inconsistent annotations on modifiers in the corpus, as shown in Figure 5.3. Basically,

modifier-caused errors are less serious, since the class and the major part of the names are correct. Among 14 instances, we find that 12 of them are acceptable.

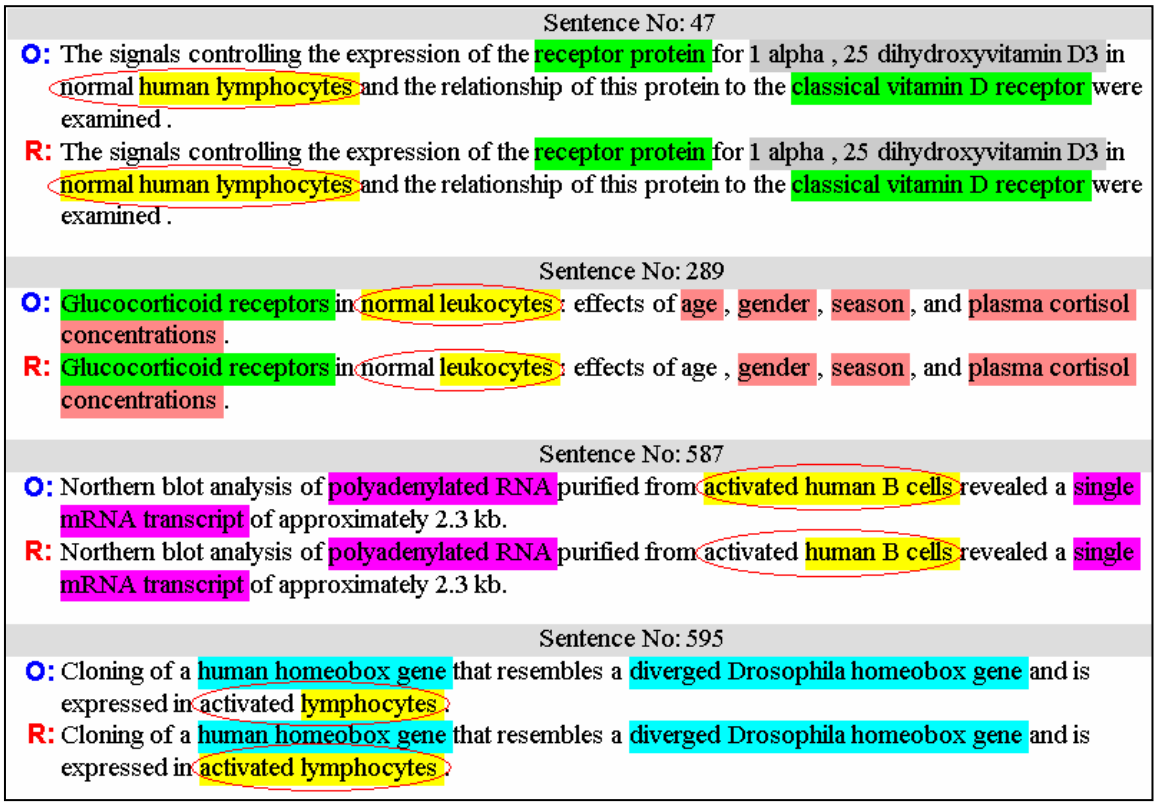


Figure 5.3: Modifier annotation inconsistency in GENIA corpus

5.3.5 Cascaded-annotation-caused Error

Cascaded-annotation-caused error accounts for the error caused by the cascaded name phenomenon. For example, the system recognizes the embedded name but not the name to the longest extent. Although we have proposed two methods for solving this problem, the inconsistency in the corpus can still affect the recognition of certain cascaded names. According to the statistics, 18 instances belong to this category. The criteria for acceptable error in this category can be different based on different understanding of usefulness of the error names. We may assert that the error is acceptable if the embedded

name is correctly recognized. Based on this criterion, we find that 11 of the 18 instances are acceptable.

5.3.6 Miscellaneous Error

This category includes all error instances that do not belong to any of the above categories. There are only 4 instances out of all the samples. None of them is acceptable.

5.3.7 Summary of Error Analysis

From the error analysis, we find that about 46% (46/100) of the errors are acceptable. It means we can reach an acceptable F-measure at about 81~82. If we have more consistent annotation and other available resources such as dictionary, it is available for further performance improvement.

Chapter 6

CONCLUSION

6.1 Conclusions

The research work presented in this thesis proposes and explores a machine learning method for biomedical name recognition. Based on detailed analysis of special characteristics in biomedical names, a rich set of features including orthographic, morphologic, part-of-speech and semantic trigger features are proposed. All these features are integrated effectively via a HMM model with back-off modeling. In addition, we also present an abbreviation recognition method and two effective cascaded name recognition methods in order to cope with the special phenomena in biomedical names. To our best knowledge, our work is the first detailed research work concerning about the cascaded name phenomenon in biomedical name recognition.

Through extensive experiments, we have following findings. First, adaptation of the part-of-speech tagger to the biomedical domain is critical for high performance biomedical name recognition. Second, the part-of speech feature, the head noun feature and the cascaded name recognition methods are very useful for biomedical name recognition. Third, the special verb trigger cannot lead to positive effect in our model. Fourth, in our model, at least 200K words are required in the training data to achieve reliable performance.

We can also find that although name recognition in biomedical domain is difficult, we can make use of the underlying information by analyzing the domain in detail. Experimental results prove that our model performs well on the GENIA corpus and outperforms previous best published system on the same training and test data. Further analysis shows that with much higher performance can be expected with a more consistently annotated corpus and a reasonable dictionary.

6.2 Future Work

From the aspect of biomedical names, further explorations can be made on the complicated constructions in the biomedical documents. One possible way is to develop effective patterns for conjunction and disjunction construction. In addition, existing public resources and databases are potential information sources that can be integrated in biomedical name recognition.

From the machine learning point of view, since names in the biomedical domain are constantly changing and there are many sub-domains that people are interested in, we should try to reduce the human work, such as annotation of large corpus, for supervised machine learning approaches. One possible solution is unsupervised or semi-supervised methods, e.g. co-training and active learning, etc., which try to get high performance by requiring the human annotation as little as possible.

6.3 Dissemination of Results

This thesis presents a coherent work on the explorations of the biomedical name recognition. The work on the analysis of biomedical names, the exploration of the feature set, and the proposal of abbreviation recognition and post-processing rule approach for cascaded name recognition methods is covered in our paper published in the *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*. [Dan et al. 2003] The work on the analysis of characteristic of biomedical names, the back-off HMM model, the further experiments on the feature set and the error analysis is covered in our paper accepted by the *Bioinformatics*. [Zhou et al. 2003] The work on the HMM-based iterative recognition approach concerning about the cascaded name recognition is to be published in the *Journal of Biomedical Informatics, Special Issue on Natural Language Processing in Biomedicine: Aims, Achievements and Challenge*. [Zhang et al. 2004]

Furthermore, a system covering all parts in this thesis is implemented and a web-demo can be accessed at http://textmining.i2r.a-star.edu.sg/NLS/demo.htm.

REFERENCES

[Bikel et al. 1999] Daniel M. Bikel, Richard Schwartz and Ralph M. Weischedel. 1999. An Algorithm that Learns What's in a Name. *Machine Learning* 34:211-231. Special Issue on Natural Language Learning.

[Borthwick et al. 1998] Andrew Borthwick, John Sterling, Eugene Agichtein and Ralph Grishman. 1998. NYU: Description of the MENE Named Entity System as Used in MUC-7. In *Proceedings of the MUC-7*.

[Borthwick 1999] Andrew Borthwick. 1999. A Maximum Entropy Approach to Named Entity Recognition. *Ph.D. Thesis. New York University*.

[Chang et al. 2002] J. T. Chang, H. Schutze and R.B. Altman. 2002. Create an Online Dictionary of Abbreviation from MEDLINE. *Journal of the American Medical Informatics Association (JAMIA)*.

[Chieu and Ng 2002] Hai-Leong Chieu and Hwee-Tou Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp. 190-196.

[Collier et al. 2000] Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. 2000. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. In *Proceedings of COLING 2000*, pages 201-207.

[Fukuda et al. 1998] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. Toward Information Extraction: Identifying protein names from biological papers. In

Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98), pages 707-718, January.

[Gaizauskas et al. 2000] Robert Gaizauskas, George Demetriou and Kevin Humphreys. Term Recognition and Classification in Biological Science Journal Articles. 2000. In *Proceedings of the Computional Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP*, pages 37-44.

[Humphreys et al. 1998] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks. 1998. University of Sheffield: Description of the LaSIE-II System as Used for MUC-7. In *Proceedings of the MUC-7*.

[Kazama et al. 2002] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and J. Tsujii. 2002. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In *Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*, pages 1-8.

[Krupka and Hausman 1998] G. R. Krupka and K. Hausman. 1998. IsoQuest Inc.: Description of the NetOwl™ Extractor System as Used for MUC-7. In *Proceedings of the MUC-7*.

[Lee et al. 2003] Ki-Joong Lee, Young-Sook Hwang and Hae-Chang Rim. 2003. Two-Phase Biomedical NE Recognition based on SVMs. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine,* pp. 33-40.

[Liu et al. 2002] H. Liu, A.R. Aronson and C. Friedman. 2002. A Study of Abbreviations in MEDLINE Abstracts. *American Medical Informatics Association Symposium.*

[Mikheev et al. 1998] A. Mikheev, C. Grover and M. Moen. 1998. Description of the LTG System Used for MUC-7. In *Proceedings of the MUC-7*.

[Miller et al. 1998] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel and Annotation Group. 1998. Algorithms that Learn to Extract Information; BBN: Description of the SIFT System as Used for MUC-7. In *Proceedings of the MUC-7*.

[Nobata et al. 1999] Chikashi Nobata, Nigel Collier, and Jun-ichi Tsujii. 1999. Automatic Term Identification and Classification in Biology Texts. In *Proceedings of the 5th NLPRS*, pages 369-374.

[Nobata et al. 2000] Chikashi Nobata, Nigel Collier, and Jun-ichi Tsujii. 2000. Comparison between tagged corpora for the named entity task. In *Proceedings of the ACL 2000 Workshop on Comparing Corpora*, pages 20-27.

[Ohta et al. 2002] T. Ohta, Y. Tateisi, J. Kim, H. Mima, and J. Tsujii. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT 2002*.

[Proux et al. 1998] Denys Proux, Francois Rechenmann, Laurent Julliard, Violaine Pillet and Bernard Jacq. 1998. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. In *Proceedings of Genome Inform Ser Workshop Genome Inform*, pages 72-80.

[Rabiner 1989] Lawrence R. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE, Vol. 77, No.2, Feb. 1989*, pp. 257~286.

[Schwartz and Hearst 2003] A.S. Schwartz and M.A. Hearst. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. In *Proceedings of the Pacific Symposium on Biocomputing 2003 (PSB 2003) Kauai*.

[Sekimizu et al. 1998] T. Sekimizu, H. Park, and J. Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. In *Proceedings of Genome Informatics*, Universal Academy Press, Inc.

[Shen et al. 2003] Dan Shen, Jie Zhang, Guodong Zhou, Jian Su and Chew-Lim Tan. 2003. Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine,* pp. 49-56.

[Takeuchi and Collier 2002] K. Takeuchi and N. Collier. 2002. Use of Support Vector Machines in Extended Named Entity Recognition. In *Proceedings of the Sixth Conference on Natural Language Learning (CONLL 2002)*, pages 119-125.

[Thomas et al. 2000] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. 2000. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing 2000 (PSB 2000)*, pages 541-551, Hawaii, January.

[Viterbi 1967] Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In *Proceedings of IEEE Transactions on Information Theory*, pp. 260-269.

[Yu et al. 1998] Shihong Yu, Shuanhu Bai and Paul Wu. 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7. In *Proceedings of the MUC-7*.

[Zhang et al. 2004] Jie Zhang, Dan Shen, Guodong Zhou, Jian Su and Chew-Lim Tan. 2004. Enhancing HMM-based Biomedical Named Entity Recognition by Studying Special Phenomena. To appear in *Journal of Biomedical Informatics, Special Issue*

on Natural Language Processing in Biomedicine: Aims, Achievements and Challenge.

[Zhou and Su 2002] Guodong Zhou and Jian Su. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pp. 473-480.

[Zhou et al. 2003] Guodong Zhou, Jie Zhang, Dan Shen, Jian Su and Chew-Lim Tan. 2003. Recognizing Names in Biomedical Texts: a Machine Learning Approach. To appear in *Bioinformatics*.

APPENDIX

The GENIA Ontology